DOCUMENT RESUME

ED 206 696 TM 810 597

AUTHOR	Peters, Eileen
TITLE	Basic Skills Improvement Policy Implementation Guide No. 3: Standards-Setting Manual. Revised Edition.
INSTITUTION	Educational Testing Service, Brookline, Mass.
SPONS AGENCY	Massachusetts State Dept. of Education, Boston.
PUB DATE	Mar 81
NOTE	66p.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Basic Skills; *Cutting Scores; Judges; Listening Skills; Mathematics; *Minimum Competency Testing; *Policy Formation; Reading Skills; *Scoring Formulas; Secondary Education; Verbal Ability; Writing Skills
IDENTIFIERS	Angoff Methods; Borderline Group Method; Contrasting Groups Method; Massachusetts; Nedelsky Method

ABSTRACT

The Massachusetts Policy on Basic Skills Improvement was formulated to help students master basic skills prior to high school graduation. The aim of this guide is to assist local districts in complying with the policy's requirements for the setting of minimal standards according to levels of achievement students are expected to reach on stated objectives. Four methods are recommended as acceptable standard setting procedures: the Angoff and Nedelsky methods, which are based upon judgments concerning the test items; and the Borderline Group and Contrasting Groups methods; which require judgments concerning the mastery status of the students taking the test. Each procedure is outlined and its appropriateness, advantages and disadvantages documented. Guidelines for selecting judges, and interpreting, using and monitoring standards are also provided. Since regulations do exist which govern standard setting and reporting, information is included on how to include parents, employers, secondary school students, public representatives, teachers and administrators in the standard setting procedure: in defining minimally acceptable performance prior to standard setting: serving as judges rating expected performance: reviewing proposed standards: and comparing standards derived from various methods. Sample student rating forms and judges recording forms are presented and the contrasting groups method paradox is discussed in the appendixes. (AEF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206696

Basic Skills Improvement Policy Implementation Guide #3

(Revised Edition)

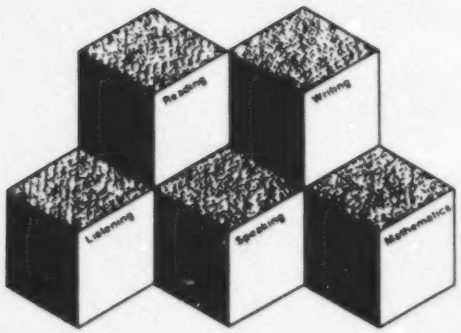
U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Standards-Setting Manual



"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. Di Bella

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 557

MASSACHUSETTS DEPARTMENT OF EDUCATION

MARCH 1981

Massachusetts Board of Education

Ann H. McHugh, Chairperson
James L. Green, Vice Chairperson

John W. Bond
James P. Doherty
Millie Clements
James R. Grande
Howard A. Greis
Charles T. Grigsby
Mary Ann Hardenbergh
Armando Martinez
Edwin M. Rossman
Donald R. Walker

Gregory R. Anrig, Commissioner of Education, Secretary

Ex. Officio Member:

Laura Clausen, Chancellor, Board of Higher Education

Bureau of Research and Assessment

Allan S. Hartman, Director
Tracy Libros
Leslie S. May
Matthew H. Towle
Betty Hancock
Cathy Ridge
Hyla Shapiro

The Massachusetts Department of Education insures equal employment/educational opportunities/and affirmative action, regardless of race, color, creed, national origin, or sex, in compliance with Title IX, or handicap, in compliance with Section 504.

Publication #12, 285-500-50-2-81-CR, approved by Alfred C. Holland, State Purchasing Agent

BASIC SKILLS IMPROVEMENT POLICY

IMPLEMENTATION GUIDE #3

(REVISED EDITION)

STANDARDS-SETTING MANUAL

Massachusetts Department of Education

March 1981

Copyright © 1981 Massachusetts Department of Education.

Some materials copyrighted by the California Department of Education are used by permission. This publication is restricted for use in the Massachusetts schools for the specific purpose of implementing the Massachusetts Basic Skills Improvement Policy and Regulations and may be duplicated by Massachusetts schools for that purpose. Circulation, distribution, and duplication for use for other purposes requires separate permission from the Massachusetts Department of Education.

ACKNOWLEDGEMENTS

The guidelines discussed in this manual are based primarily on Michael Zieky's and Samuel Livingston's description of standards-setting procedures in Educational Testing Services' Manual for Setting Standards on the Basic Skills Assessment Tests, (Princeton, New Jersey, 1977). Materials on interpreting and using standards are derived from the Handbook for Proficiency Assessment, (Sacramento, California: California State Department of Education, 1979).

The present manual is a revised, updated version of Implementation Guide #3, developed for the Bureau of Research and Assessment by the ETS Northeastern Regional Office. This revised version includes many new materials, including additional information on the contrasting-groups method developed by the RMC Research Corporation. The final responsibility for the contents of this manual, however, rests solely with the Massachusetts Department of Education.

Special Note: Please be reminded that all numerical data included in this manual are cited solely for the purposes of illustrating how standards-setting methods work. These examples should not influence in any way the local performance standards that are to be determined independently by each school district as called for in the state Regulations for Implementation of the Policy on Basic Skills Improvement.

Contents

		page
I.	OVERVIEW	1
	Policy Requirements	1
	What is a Standard?	1
	Why Set Standards?	1
	Unacceptable Methods	2
	Acceptable Methods	3
II.	SOCIAL AND POLITICAL ISSUES RELATED TO STANDARDS-SETTING	5
III.	ITEM REVIEW METHODS OF SETTING STANDARDS	7
	Selecting Judges	7
	Defining Minimally Acceptable Performance	7
	Collecting and Combining Judgments	8
	Nedelsky's Method	9
	Reminder List for the Nedelsky Method	13
	Advantages/Disadvantages of the Nedelsky Method	14
	Angoff's Method	15
	Reminder List for the Angoff Method	18
	Advantages/Disadvantages of the Angoff Method	19
IV.	STANDARDS-SETTING METHODS BASED ON JUDGMENTS OF STUDENTS' MASTERY STATUS	20
	The Borderline-Group Method	21
	Reminder List for the Borderline-Group Method	24
	Advantages/Disadvantages of the Borderline Group Method	25
	The Contrasting-Groups Method	26
	Reminder List for the Contrasting-Groups Method	30
	Advantages/Disadvantages of the Contrasting-Groups Method	31
V.	INTERPRETING AND USING STANDARDS	32
	Two Types of Errors of Classification	32
	Determing if the Cutting Score is Reasonable	34
	Adjusting the Cutting Score	36
	An Alternative to a Single Cutting Score	38
	Using Multiple Methods to Set Cutting Scores	39
	Monitoring Standards	39
VI.	REGULATIONS GOVERNING STANDARDS-SETTING AND REPORTING	41
	How Does Public Participation Fit into the the Standards-Setting Process?	41
	Suggestions for Public Participation	42
	What Should be Reported?	43
APPENDICES		

I. OVERVIEW

Policy Requirements

The purpose of this manual is to assist local districts in complying with the requirements of the Massachusetts Policy on Basic Skills Improvement, whose principal goal is to help students master basic skills prior to high school graduation. To accomplish this goal, the policy requires schools to develop effective basic skills programs, a component of which is minimal performance standards. For each skill area (reading, writing, mathematics, listening, and speaking), minimal standards are set according to the level of achievement students should reach on stated objectives.

What is a Standard?

Such terms as "passing score," "cutting score," "cutoff point," or "mastery level" are often used to refer to a specified level of achievement, or a standard. We might decide, for example, that a student has mastered the objectives in question if the student gets at least 43 of the 70 test questions (measuring those objectives) correct. By making such a decision, we would be setting the standard at 43. A standard is the lower limit of the range of acceptable performance on a given test or other measure. In our example, 43 is the lowest score a student can get and still be considered a master of the skills measured by the test.

Why Set Standards?

Minimal performance standards are used to identify students who may require further instruction in the various skill areas. There would be no need to set such standards if students who have mastered the content measured by a test would always answer all the test questions correctly and if students who have not mastered the content would all get scores of zero. But in the real world of teaching and testing, we rarely get such clear-cut results. Obviously, students with "low" scores are more likely to require instructional help than students with "high" scores. Our task

is to decide the lowest score a student can get and still be considered a master--one who appears not to need further instruction.

State of the Art: Unacceptable Methods

It is extremely important to realize that all methods of setting standards depend on subjective judgment at some point in their application. There is no good way of setting standards just by plugging numbers into a formula. Although the subjectivity of standards-setting is inescapable, some methods of collecting judgments are more defensible than others.

Traditional approach. The method with the least theoretical justification is the one most often used in practice: the arbitrary choice of a fixed percent-correct as the standard, without regard to the length, difficulty, or importance of the test. An example is the traditional reliance on 70 percent correct as "passing," without reference to the test or the population.

Judgment not based on data. Although the process may become extremely elaborate, involving broad-based committees, questionnaire surveys, and the like, the decision is an arbitrary one. It is arbitrary because it does not take into account test variables (such as specifications and item difficulty) or performance on some criterion measure (e.g., job tasks).

Quota-setting. Using this method, students are ranked on some performance measure such as a test score. Then they are selected (from the highest score down on the bottom score up) to participate in a program. The number selected depends upon the number that can be accommodaced--the resources available. If, for example, a school district decided it could provide additional reading resources for only 15 percent of its students, the reading test score that separated the lowest 15 percent from the group as a whole would be designated the standard.

Acceptable methods

Fortunately, there are acceptable standards-setting methods--methods that have theoretical justification and yet do not require too much hard-to-get data. These methods can be applied without specialized training, and the calculations they require are not difficult.

As noted above, all methods of setting standards require some type of judgment. The two main types of judgments that can be used to set test performance standards are (1) judgments about the questions in the test and (2) judgments about the mastery status of a sample of students taking the test.

This manual describes four methods: two for each type of judgement.

Judgments about test questions -

1) Nedelsky's Method

2) Angoff's Method

Judgments about students' mastery status -

1) The Borderline-Group Method

2) The Contrasting-Groups Method

These methods are appropriate for setting standards in reading, listening, mathematics, and, in some cases, writing. Because writing is a somewhat special case, however, writing standards are treated fully in Implementation Guide #2 (Revised Edition): Assessing Writing Skills.

Some cautions. A dangerous misconception about cutting scores is that they are absolute and precise indicators of who has achieved mastery. Remember that tests are not perfect measuring instruments, just as students are not perfect test takers and cutting scores are not set by perfect judges. Given these sources of error, there is little assurance that a student who scores just above the standard is superior to a student who scores just below it. A

standard, carefully set, simply assures us that on the average we will make fewer mistakes using that score than using any other score. No single procedure can assure us that we will not make any mistakes. A test score is, after all, only a piece of information about a student. For an overall picture of a student's achievement, a test score should be interpreted in light of other information including courses taken, classroom performance, teacher judgments, and other test data. (These issues are treated further on pages 32 - 40 of this manual.)

II. SOCIAL AND POLITICAL ISSUES RELATED TO STANDARDS-SETTING

The setting of standards involves social and political issues as well as measurement issues. Even if highly sophisticated statistical procedures are performed flawlessly, the standards-setting process will be deficient unless the issues discussed below are considered.

1. Standards Should Be Viewed as the Results of a Process

The process used to set standards will have a great impact on the acceptability of the standards that are set. The judges used to help set standards should include representatives of both sexes as well as racial, ethnic, and cultural minorities in the community. In addition, representatives of such interested groups as school committees, local educational associations, and parent-teacher associations should be included. If students from more than one school take the test, be sure to involve staff from each of the schools in the process of setting standards. Suggestions for community involvement are included in Section VI of this manual.

The fact that standards will be set and the ways in which they will be set should be well publicized. Members of the community should have the opportunity to provide input at meetings established for that purpose. The more open the process, the more likely it is that the standards will be widely accepted and supported.

2. Standards Should Be Viewed as Tentative Estimates

It is unlikely that a school district's first attempt at setting standards will work as well as subsequent ones. The experience gained in the first round should help improve the following attempts.

In addition, the educational needs and experience of a school district are likely to change over time, which may require changes in standards. Suggestions for reviewing standards are presented in section V of this manual.

3. Standards Should Not Be Overgeneralized

The results of the application of standards should not be used to measure teacher or program effectiveness. There are too many variables, other than teacher effectiveness, that affect student performance to allow such use of the test results in other than controlled, experimental settings.

4. Setting Standards Should Not Be Hurried

It should become clear from the contents of this manual that attempting to set standards is a major task. The person given responsibility for managing this task will need time to plan and carry out his or her work. If you do not spend the necessary time, it is not likely that standards-setting will be done appropriately in your district.

After reading this manual and selecting the method(s) you are going to use to set standards, it is a good idea to try a "dry run" or pilot standards-setting exercise with a small number of judges and/or a small sample of students. Such a process should help polish your skills and increase the chances that the full-scale standards-setting will run smoothly.

III. ITEM REVIEW METHODS OF SETTING STANDARDS

The two methods described in this section of the manual are based on judgments of what certain types of students should be able to do. The judgments of what students should be able to do are made in terms of the specific content of a test, one question at a time. Judgments about individual test questions are then combined to produce a standard for the full test. Each method consists of four basic procedures:

1. Select judges.
2. Define "minimally acceptable performance."
3. Collect judgments.
4. Combine the judgments to set a standard.

Procedures 1 and 2 are the same for both methods; procedures 3 and 4 differ.

Procedure 1 -- Selecting Judges. The judges should be a group of persons who are qualified to decide what the students should be able to do. Teachers at the appropriate grade levels are an obvious choice and should always be involved in the task, but you should also consider using local people such as employers, school committee members, and community representatives. The more judges, the better their combined judgment will represent the judgment of all persons who could have been judges. You should use at least five judges, and more if you can get them. However, the process becomes cumbersome with too many judges. Twenty would be a reasonable upper limit, for practical reasons.

Procedure 2 -- Defining Minimally Acceptable Performance. As a first step, have the judges review your basic skills objectives, and make sure they understand what is measured by the test. Then ask them to describe, in their own words, the lowest level of performance (in terms of the objectives) they would consider acceptable.

The judges may find it convenient to describe the performance of specific students with whom they have worked and whose performance they consider minimally acceptable.

Encourage discussion and use lots of examples. Ask questions such as: "Should minimally acceptable performance require the students to be able to get specific details from a newspaper article? Should it require students to distinguish statements of fact from statements of opinion? Should the student be able to state the main idea of a Reader's Digest article? How about an article from Scientific American?"

Be prepared to spend two hours or so reaching and recording a definition of minimally acceptable performance for your district. The definition should be sufficiently clear to permit easy communication among the people contributing to the task of standards-setting. Furthermore, the definition should be understandable to the lay public. It is best to avoid educational jargon and esoteric terminology. Be sure to record the definition that you arrive at because there is certain to be widespread interest in it.

Procedures 3 and 4 -- Collecting and Combining the Judgments. These procedures distinguish one method from the other. We will present two methods that are usually identified by the names of the authors who first suggested them in published writings: Nedelsky and Angoff.

Nedelsky's Method

How does the Nedelsky method work?

The method requires a group of judges to review a test and make a decision for each question on the number of wrong alternatives a minimally competent student would be expected to identify as obviously wrong. The number of alternatives remaining constitutes the set from which the student is expected to guess. The procedure can be used only with multiple-choice tests.

How should you prepare for the item review session?

1. Schedule a work session of about three hours for discussing your objectives and reviewing an average length test (60 items). You should plan two three-hour sessions to conduct the preliminary discussions and review the state reading and mathematics tests, and a one-hour session to cover the state listening test. Other tests may require more or less time, depending on the number of test questions.

2. Obtain enough copies of each test to be able to distribute one to each judge. (An overhead projector can be used, but it is better if each judge has a copy.)

3. Make enough copies of the Judge's Recording Form: Nedelsky Method for each judge. (See Appendix A.)

4. Make copies of the district's basic skills objectives for each judge.

5. Arrange for a comfortable meeting room with sufficient table or desk space.

6. Obtain a copy of the test manual which includes the test answers and information on item difficulty (p-values).

7. Have a calculator available.

8. Have tapes available for listening tests.

What steps should you follow in conducting tne item review session?

1. Give each judge a copy of the test and the recording form.

2. Review the purpose of the tests and the standards-setting exercise, including a presentation and discussion of the district's basic skills objectives.

3. Have the judges look at the first test question and use the following format:

   A. Focus the judges' attention on the first wrong answer. Ask how many thought the minimally competent student should be able to eliminate it and how many did not think so.

   B. If the judges are not unanimous, ask for one judge who identified that wrong answer to explain why. Then ask for one judge who did not identify it to explain why not. Do not try to reach agreement; just allow each point of view to be heard. The judges may or may not be swayed by the comments of their colleagues.

   C. Go on to the next wrong answer.

4. After the question has been discussed in this way, ask the judges to independently circle, on their recording forms, the number of wrong answers a minimally competent student could identify as wrong.

5. After the judges have worked through the first few questions and fully understand the process, you have two alternatives. Either continue using the same full group discussion format until all the questions in the test have been completed, or have the judges work independently on the remaining test questions.

6. If judges work independently, periodically have them raise their hands on each wrong alternative. A difficulty index for each item can be calculated quickly by using the probabilities on the recording forms. If the judges' expectations consistently exceed the average performance of a comparable population of students (see test manual or local data), the standard most likely will be unrealistically high.

How should judgments be combined and standards set?

1. Convert each Judge's Recording Form to a table of probabilities. Conceptually, the way to do this is to assume that the student will guess at random from among the answers he or she does not know to be wrong. For example, if a question has four possible answers and the student can rule out one of them, the student will have three answers left to choose from.[1] This gives the student a one-third chance of guessing correctly. Operationally, the process is a simple clerical one. Take each Judge's Recording Form. For each question, in the column headed P write:

   ".25" (if the judge circled 0)

   ".33" (if the judge circled 1)

   ".50" (if the judge circled 2)

   "1.00" (if the judge circled 3)

2. Use a hand calculator to add up the numbers in column P on each form. Write the sum in the box labeled SUM.

3. Now just calculate the average of the sums. Simply add all the sums and divide by the number of sums. The number of sums should, of course, be equal to the number of judges.

The result is an estimate of the standard for the test.

The following example shows how recording forms completed by three judges are used to set a standard for a hypothetical test consisting of ten questions, each with four possible answers.

---

[1] Other probabilities are as follows:

Probabilities for five-option test questions		Probabilities for three-option test questions	
.20	(if the judge circled 0)		
.25	(if the judge circled 1)	.33	(if the judge circled 0)
.33	(if the judge circled 2)	.50	(if the judge circled 1)
.50	(if the judge circled 3)	1.00	(if the judge circled 2)
1.00	(if the judge circled 4)		

Nedelsky's Method

Sample Standards Based on Three Sets of Judgments on a Ten Item Test

Question	Judge A Circle Numbers of Wrong Answers	P
1	0 1 2 (3)	1.00
2	0 1 (2) 3	.50
3	0 1 (2) 3	.50
4	0 1 2 (3)	1.00
5	0 (1) 2 3	.33
6	0 (1) 2 3	.33
7	0 1 (2) 3	.50
8	0 (1) 2 3	.33
9	(0) 1 2 3	.25
10	0 (1) 2 3	.33
		SUM 5.07

Question Number	Judge B Circle Numbers of Wrong Answers	P
1	0 1 2 (3)	1.00
2	0 1 2 (3)	1.00
3	0 1 (2) 3	.50
4	0 1 2 (3)	1.00
5	0 (1) 2 3	.33
6	0 (1) 2 3	.33
7	0 (1) 2 3	.33
8	(0) 1 2 3	.25
9	0 (1) 2 3	.33
10	(0) 1 2 3	.25
		SUM 5.32

Question Number	Judge C Circle Numbers of Wrong Answers	P
1	0 1 2 (3)	1.00
2	0 1 2 (3)	1.00
3	0 1 (2) 3	.50
4	0 1 (2) 3	.50
5	0 1 (2) 3	.50
6	0 (1) 2 3	.33
7	0 1 (2) 3	.50
8	(0) 1 2 3	.25
9	(0) 1 2 3	.25
10	0 (1) 2 3	.33
		SUM 5.16

How to Compute the Standard

1. The sums of the numbers in each column P are 5.07, 5.32, and 5.16.

2. The average of the sums equals 5.18.

$$\begin{array}{r} 5.07 \\ 5.32 \\ 5.16 \\ \hline 15.55 \end{array} \qquad 3\overline{)15.55} \quad 5.18$$

3. The estimated standard, 5.18, falls between 5 and 6. A student scoring 5 or less would fall below the standard. A student scoring 6 or more would be above the standard.

Reminder List for the Nedelsky Method

In carrying out this standards-setting procedure, be sure to:

- Involve a number of teachers (at least five or six) at each grade level in each skill area.
- Schedule sufficient time for the item review session.
- Have recording forms that match both the number of items in the test and the item format (3-option, 4-option, or 5-option items).
- Circulate the district's basic skills objectives to the judges.
- Clearly describe the standards-setting tasks and rationale.
- Watch out for judge fatigue in the review session, and provide for breaks.
- Allow for discussion of different viewpoints.
- Double-check all calculations.

NEDELSKY METHOD

Advantages

1. Procedure is independent of numbers of students taking the test. Cutting score can be calculated for very small groups of students.
2. Procedure can accommodate participation by a broad cross-section of community members as judges (e.g., teachers administrators, parents, students).
3. Procedure is based on close scrutiny of items in the instrument.
4. Procedure losely parallels decision processes of test takers; each item must be rejected or accepted.

Disadvantages

1. Procedure can be used only with multiple-choice tests.
2. Actual test performance is not taken into account.
3. Cutting score can be too high or low when there are a disproportionate number of judges with the same interest or bias.
4. A very limited number of probabilities can be assigned.

Angoff's Method

How does the Angoff method work?

The Angoff method, a probability estimation method, is similar in many ways to the Nedelsky method but is more flexible in the assignment of probabilities. In addition, the method can be used for both multiple-choice and open-ended questions, while the Nedelsky method is appropriate only for multiple-choice tests.

The basic procedure requires a judge to examine each item in the test and directly decide the probability that a minimally competent student would answer the item correctly. Each test question is judged in its entirety, in contrast to the Nedelsky method which requires separate judgments for each response option.

How should you prepare for the item review session?

1. Schedule a work session of about three hours for reviewing an average length test (60 items). You should plan two three-hour sessions to conduct the preliminary discussions and review the state reading and mathematics tests, and a one-hour session to cover the state listening tests. Other tests may require more or less time, depending on the number of test questions.

2. Obtain enough copies of each test to be able to distribute one to each judge. (An overhead projector can be used, but it is better if each judge has a copy.)

3. Make enough copies of the Judge's Recording Form: Angoff Method for each judge. (See Appendix B.)

4. Make copies of the district's basic skills objectives for each judge.

5. Arrange for a comfortable meeting room with sufficient table or desk space.

6. Obtain a copy of the test manual, which includes the test answers and information on item difficulty (p-values).

7. Have a calculator available.

8. Have tapes available for listening tests.

Collecting the Judgments

1. Give each judge a copy of the test and the recording form.

2. Have the judges look at the first question and state the probability that a minimally competent student would answer the question correctly. This task may require some time to explain. If the judges are not comfortable dealing with probabilities, ask them to think of a group of 100 minimally competent students and state how many of those students would be expected to answer the question correctly. Obviously, the easier the question, the higher the probability will be. The probability must be between .00 and 1.00. If the questions are multiple-choice, the probability generally should not be lower than the chance of guessing the correct answer by blind luck. For a 4-option question, the probability should not be below .25 (1 chance in 4 of guessing correctly). For a 5-option item, the probability should not be lower than .20. For a three-option item, the probability should not be lower than .33.

3. Have each judge announce his or her choice of a probability that a minimally competent student would answer the first question correctly. Write these numbers on a blackboard or a large sheet of paper so that all the judges can see them. Then ask the judges who stated the largest and smallest numbers to explain the reasons for their choices. Then tell the judges they may change their choices if they want, but not to announce them. Instead, they should simply write their revised choices on the Judge's Recording Form.

4. After you are certain that the judges understand the process, they can either work independently on the remaining test questions, or continue the group review until all test questions are completed.

5. If judges work independently, monitor their work periodically. If the judges' expectations regarding difficulty consistently exceed the average performance of a comparable population of students (see test manual or local data), the standard will most likely be unrealistically high.

Combining the Judgments

What steps should you follow in setting the standard?

1. Collect the forms from the judges.

2. Use a hand calculator to add up the probabilities on each judge's form. Write the sum in the box labeled SUM.

3. Now calculate the average of the sums. Simply add all the sums and divide by the number of sums (the number of judges).

The following example shows how recording forms completed by three judges are used to set a standard.

Angoff's Method

Sample Standards Based on Three Sets of Judgments on a Ten Item Test

Judge A	
Question Number	Estimated Probability
1	1.00
2	.90
3	.80
4	.70
5	.35
6	.45
7	.25
8	.30
9	.25
10	.25
	SUM 5.25

Judge B	
Question Number	Estimated Probability
1	1.00
2	.85
3	.85
4	.70
5	.35
6	.40
7	.25
8	.30
9	.25
10	.25
	SUM 5.20

Judge C	
Question Number	Estimated Probability
1	.90
2	.80
3	.80
4	.65
5	.40
6	.45
7	.35
8	.35
9	.30
10	.25
	SUM 5.30

How to Compute the Standard

1. The sums of the probabilities assigned by the judges are 5.25, 5.20, and 5.30.

2. The average of the sums equals 5.25.

$$\begin{array}{r} 5.25 \\ 5.20 \\ \underline{5.30} \\ 15.75 \end{array} \qquad 3\overline{)15.75} \quad 5.25$$

3. The standard, 5.25, falls between 5 and 6. A student scoring 5 or less would fall below the standard. A student scoring 6 or more would be above the standard.

Reminder List for the Angoff Method

In carrying out this standards-setting procedure, be sure to:

- Involve a number of teachers (at least five or six) at each grade level in each skill area.
- Schedule sufficient time for the item review session.
- Have recording forms that match the number of items in the test.
- Circulate the district's basic skills objectives to the judges.
- Clearly describe the standards-setting tasks and rationale.
- Watch out for judge fatigue in the review session, and provide for breaks.
- Allow for discussion of different viewpoints.
- Double-check all calculations.

ANGOFF METHOD

Advantages

1. Procedure is independent of numbers of students taking the test. Cutting score can be calculated for very small groups of students.
2. Procedure can accommodate participation by a broad cross-section of community members. (e.g., teachers, administrators, parents, students)
3. Procedure is based on scrutiny of each item on the test.
4. Allows more flexibility in assigning probabilities than Nedelsky Method.
5. Procedure is not limited to multiple-choice items.

Disadvantages

1. Procedure does not take actual test performance into account.
2. Passing score can be too high or low when there are a disproportionate number of judges with the same interest or bias.
3. Procedures may be difficult for some judges uncomfortable with abstract assignment of probabilities.

IV. STANDARDS-SETTING METHODS BASED ON JUDGMENTS OF STUDENTS' MASTERY STATUS

Two methods for setting standards which depend on students' actual test performance--the borderline method and the contrasting-groups method--are presented below. Both methods require judgments about the students who take the test rather than judgments about the test items themselves.

The methods assume that distinguishable groups of students in the school population can be identified. That is, based on performance in a given area such as mathematics, students can be classified meaningfully as masters, non-masters, or borderline-masters. It is further assumed that a teacher who has instructed a student in the skill area long enough to be reasonably certain of his or her level of achievement is the person who can make the best a priori and independent judgment of the student's mastery status. Several other conditions are essential for valid teacher judgment.

- A judgment must be made on the basis of the student's achievement in the basic skills, not on the student's attitude, attendance or other behavioral issues. A student's mastery of a basic skill often may be difficult to isolate from these variables. The scores of students with attitudinal or other non-academic problems, as well as those of recent transfer students, should not be used in calculating the standard.

- A judgment should be made by a teacher who knows the student's skills relative to the test content and the level of test difficulty. Generally, mathematics teachers should judge students for the mathematics test and English teachers should judge reading and writing. However, even mathematics and English instructors who are teaching higher level courses (such as algebra and literature) may sometimes have difficulty focusing on the basic level of mastery. Depending on the course structure and content, others (such as social studies teachers) may be able to make adequate classifications.

- The district's objectives and definition of mastery or minimal competence in each skill area should be made clear to the teachers. The purposes of the test and the reason for collecting their judgments should also be explained.

- Judgments must be based on a student's present achievement. Judgments about a student's performance that are even a few months old may not be accurate indicators of the student's subsequent mastery.

The Borderline-Group Method

How does the borderline-group method work?

In each area (such as secondary level reading) for which standards are being set, teachers identify those students whose skills are judged to be not quite adequate, but not quite inadequate. They are "borderline-masters" of the basic skills objectives in question (such as secondary level reading objectives). After these students are identified, they are administered a test designed to measure basic skills in the area of interest. The median test score obtained by borderline students is then calculated and chosen as an estimate of the minimal standard.

How is this method carried out?

1. Determine the subject areas and grade levels for which you will set standards using this method, for example, secondary (8th grade) level reading, listening, and mathematics.

2. Estimate the number of students who are likely to be borderline-masters of each skill area. If, for example, you are setting reading standards at the 8th grade level, estimate the number of 8th graders whom you guess might fall into the borderline category for reading. If you expect sizable groups, roughly 100 borderline-masters of each skill area, proceed to the next step.

3. Identify the students who are to be tested and rated by their teachers.

4. Identify the appropriate teachers. Remember, it is crucial that raters be familiar with students' performance levels. Select mathamatics teachers to judge students' mastery of mathematics objectives, English teachers to judge mastery of reading objectives, and so on. Also remember that judgments must be current. If you are setting 8th grade mathematics standards, for example, select only 8th grade mathematics teachers to judge students' math skills.

5. Circulate the appropriate set of basic skills objectives to each teacher you have identified. Distribute class rosters and instructions for rating students. (A sample set of instructions appear as Appendix C to this manual.)

6. Ask each teacher to submit a list of students whose performances are so close to the borderline between acceptable and unacceptable that they cannot be classified into either group. Ratings must be completed and lists returned before students are tested.

7. Administer the basic skills tests for which standards are being set.

8. Compile teachers' lists to determine the areas (such as secondary level reading) for which you have approximately 100 students rated as borderline-masters. As a rule of thumb, do not drop below 90 borderline-masters. Although there is nothing magical about the number 90, it represents a reasonable lower limit for obtaining consistent results.

How is the standard calculated?

1. For each grade level and subject area of interest, compile a list of test scores obtained by the 90 or more students classified as borderline. (Do not include the scores of masters, non-masters or students who could not be rated.)

2. Rank borderline students (from high to low) on the basis of their test scores. (You will have separate rank orders for each test you administer.)

3. Count the number of students who were judged borderline and the number for whom you have test scores. (These two numbers should be identical.)

4. Identify the median score by counting down the list of scores until you reach the middle score. If the number of scores is even, average the two middle scores. (See the example on the following page.)

BORDERLINE-GROUP METHOD

Sample Score Distribution

Score		Number of Borderline-Masters at Score Point
76		2
75		2
74		1
73		2
72		4
71		5
70		4
69		7
68		10
67		9
66		11
65		14
64	<-Median or Middle Score	15
63		12
62		11
61		13
60		12
59		6
58		5
57		3
56		4
55		2
54		3
53		2
52		1

Reminder List for Borderline-Group Method

- Do not use this method unless you have at least 90 students classified as borderline.
- Remember that it is crucial that the judges be familiar with the students' levels of performance. Be very careful to collect judgments of mastery levels in mathematics from students' math teachers, and so on.
- Be sure that teachers' ratings pertain only to achievement and are not based on attitudinal or behavioral characteristics.
- Be sure that the borderline category reflects skill proficiency and is not used to identify students whose work the teacher is not able to judge.
- Collect teachers' ratings before tests are administered.
- Develop separate rank-order lists for each level/skill area for which you are setting standards (such as secondary/listening).
- Double-check test scores, rank-order lists, and calculations.

BORDERLINE-GROUP METHOD

Advantages

1. Procedure is based on actual student performance data.
2. Judgments are made by teachers who are knowledgeable about the subject matter and students.
3. Judgment collection may be quicker than for item review procedures.
4. Median score is simple to calculate.

Disadvantages

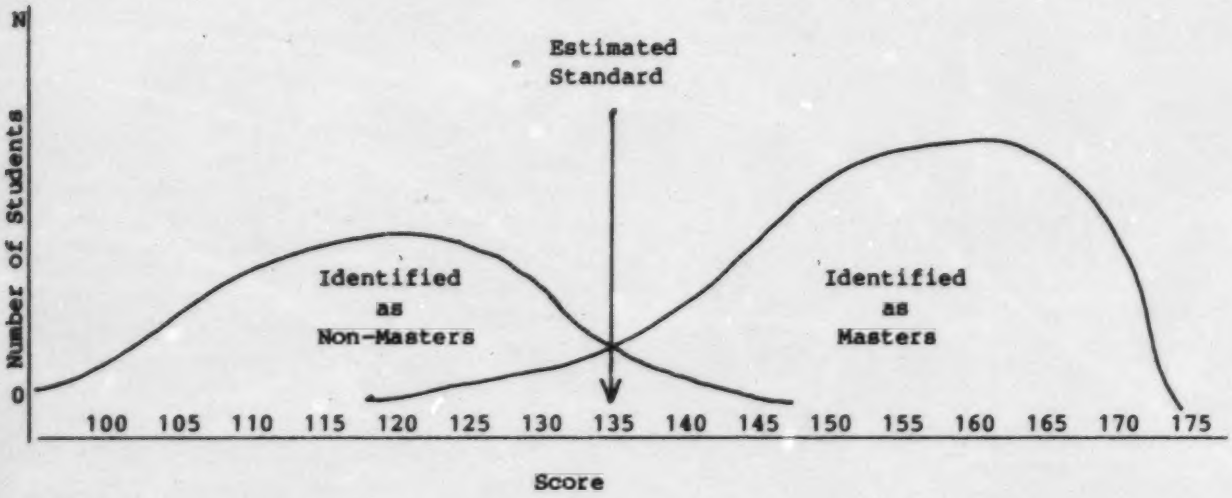
1. For reliable calculation, approximately 100 borderline students should be identified.
2. Although the public can be involved reviewing the standard and the objectives, only teachers can judge students' mastery levels.
3. At the secondary level, judgments may be unobtainable for students who are not taking a course in the subject matter being tested.
4. Testing cannot occur until teachers have become familiar with students' abilities.
5. Misidentification is possible if definitions of mastery and non-mastery are vague and unevenly applied.

The Contrasting-Groups Method

How does the contrasing-groups method work?

For the contrasting-groups method, teachers identify students whom they are sure are either definite masters or definite non-masters of the objectives measured by the test on which a standard is being set. Students are then tested, and the scores obtained by those identified as masters are compared with the scores achieved by the non-masters. As shown below, the standard is generally set at the point where the two groups of scores intersect. Before making this decision, however, you should read the material presented in Section V of this manual.

Exhibit 1



For consistent results, you need about 100 students in the smaller of the two groups (generally those classified as non-masters). For example, if you are setting mathematics standards for grade 8, you should not use this method unless teachers have rated approximately 100 8th graders as non-masters of the math skills you are assessing. As a rule of thumb, do not drop below 90 students. Although there is nothing magical about the number 90, it represents a reasonable lower limit for obtaining consistent results.

How is this method carried out?

1. Determine the subject areas and levels for which you will set standards using this method, for example, secondary level listening.

2. Estimate the number of students who are likely to be non-masters of each skill area. If, for example, you are setting reading standards at the 8th grade level, estimate the number of 8th graders whom you guess might be classified as non-masters of the secondary level reading objectives. If you expect sizable groups--roughly 100 non-masters of each skill area--proceed to the next step. It is extremely important, however, not to allow the method's assumptions governing numbers of students to influence teachers' ratings.

3. Identify the students who are to be tested and rated by their teachers.

4. Identify the appropriate teachers. Remember it is crucial that raters be familiar with students' performance levels. Select mathematics teachers to judge students' mastery of mathematics objectives, English teachers to judge mastery of reading objectives, and so on. Also remember that judgments must be current. If you are setting 8th grade mathematics standards, for example, select only 8th grade mathematics teachers to judge students' math skills.

5. Circulate the appropriate set of basic skills objectives to each teacher who will rate students. Also distribute instructions for rating students (See Appendix D) and class rosters, or, if you prefer, teacher rating forms (See Appendix E).

6. Schedule a meeting for teachers who will rate students. At this meeting, explain procedures for completing and returning rating forms, and allow enough time for teachers to become familiar with the distinction between "mastery" and "non-mastery" of your district's basic skills objectives.

7. Ask teachers to rate students whom they know to be masters and non-masters of the skills in question. They can use a rating form (See Appendix E) or simply mark class rosters, for example, with an "M" next to students who are clearly masters and an "S" next to those who are clearly in need of supplemental instruction. Ratings must be completed and returned before students are tested.

8. Administer the basic skills tests for which standards are being set.

9. Compile teachers' ratings to determine the areas (such as secondary level reading) for which you have approximately 100 students identified as non-masters (or masters if the mastery group is the smaller of the two).

How to calculate the standard.[1]

Organize the data

1. Get lists of test scores and teacher judgments for each skill area. (Separate rank order lists of masters and non-masters may be easiest to work with.)

2. Delete or cross out test scores obtained by students who were not judged or who were judged as "uncertain" for that skill.

3. If the test scores are not rank-ordered, simply inspect the results to spot the highest score obtained by any student (master or non-master). Next, determine the lowest score.

4. Starting with the highest obtained score, list all consecutive scores in the left hand column of a blank piece of paper. Conclude with the lowest score obtained by any student tested (master or non-master). See the table on the following page. In this hypothetical case, the highest score is 85 and the lowest is 33. (Column 1 of the table.)

5. Now starting with students rated as masters, count and record the number of students receiving each score. (See column 2 of the table on the following page.) Then record the number of non-masters obtaining each score.

Construct cummulative frequency distributions

1. Start with the masters. For each score, record the number of masters who received a score lower than the one listed in the left hand column. In the table, for example, there are 278 masters, and only 2 received the top score of 85. (See column 3 of the table.) Therefore, 276 students (278 minus 2) received a score lower then 85. Similarly, if 8 masters obtained scores of 83, the number "below score" would be 268 (278 minus 8 who received 83, minus 2 who received 85).

2. Now turn to the non-masters. For each test score, record the number of non-masters who obtained either that score or a higher score. In the table, 2 non-masters obtained scores of 74. Thus, the number 2 is recorded in column 5, headed "number at or above score." The number is carried down the table. For example, although no non-master received a test score of 73, 2 students (scoring 74) received scores "at or above" score 73.

Calculate possible classification errors

1. For each test score, add the number of masters who actually scored lower and the number of non-masters who were at or above score. (See column 6 of the table.) For example, for score 71, there are 144 possible classification errors (140 masters who scored below 71, plus 4 non-masters who scored at or above 71).

---

[1] A paradox inherent in the contrasting-groups method, and a solution to this problem, appears as Appendix F to this manual.

Estimate the standard

1. A reasonable estimate for the standard is the score at which possible classification errors are minimized. Thus, for the hypothetical data in the table, score 58 (with 65 possible classification errors) is a reasonable estimate of the standard. Before deciding on a standard, however, you should read pages 32 - 40 of this manual.

Number of Students Possibly Misclassified at Selected Score Points

	Rated Masters by Teachers (N=278)		Rated Non-Masters by Teachers (N=107)		
Score	Number at Score Point	Number Below Score	Number at Score Point	Number at or Above Score	Possible Classification Errors
85	2	276	0	0	276
83	8	268	0	0	268
81	4	264	0	0	264
80	9	255	0	0	255
78	22	233	0	0	233
76	13	220	0	0	220
75	11	209	0	0	209
74	19	190	2	2	192
73	26	164	0	2	164
72	2	162	0	2	162
71	22	140	2	4	144
70	2	138	2	6	144
68	29	109	8	14	123
67	1	108	0	14	122
66	21	87	1	15	102
65	8	79	2	17	96
64	18	61	4	21	82
63	6	55	4	25	80
62	11	44	6	31	75
61	4	40	2	33	73
59	12	28	9	42	70
58	6	22	1	43	65
57	5	17	7	50	67
55	2	15	4	54	69
54	3	12	8	62	74
52	3	9	8	70	79
50	6	3	5	75	78
49	0	3	5	80	83
47	1	2	5	85	87
45	0	2	3	88	90
43	1	1	6	94	95
41	0	1	5	99	100
39	0	1	5	104	105
33	1	0	3	107	107

Reminder List for Contrasting-Groups Method

- Except for corroborative purposes, do not use this method unless you have at least 90 students classified as non-masters of the objectives measured by each test on which you will set standards. If, for example, you are setting reading standards at the 8th grade level, you will need a minimum of 90 8th graders rated as non-masters of the secondary reading objectives. Do not, however, allow this requirement to influence teachers' ratings.

- Schedule the teacher rating session after the teachers know the students and before test results are reported to the teachers.

- Circulate the appropriate set of basic skills objectives to each teacher who will rate students. Also distribute class rosters or rating forms and instructions for rating students. (Appendix E.)

- Have each student rated by a teacher who now teaches that student in that skill area.

- Be sure that teachers' ratings pertain only to achievement and are not based on attitudinal or behavioral characteristics.

CONTRASTING-GROUPS METHOD

Advantages

1. Procedure is based on actual student performance data.
2. Judgment is made by teachers who are knowledgeable about subject matter and students.
3. Method is psychometrically cleaner--errors of classification can be visualized.
4. Judgment collection is quicker than for item-review procedures.

Disadvantages

1. For reliable calculations, approximately 100 students are required for the smallest of the two mastery groups.
2. Direct input is missing for all affected constituencies except teachers of students being tested.
3. Judgments may be unobtainable for students who are not taking a course in the subject matter being tested.
4. Testing must occur later in the year than with item-review procedures to allow teachers time to determine students' abilities.
5. Misidentification is possible if definition of mastery and non-mastery are vague or unevenly applied by teachers. (Teacher judgments might not be made with respect to test or district's definition of minimal competence.)

V. INTERPRETING AND USING STANDARDS

A student's score on a test is not a perfect indication of the student's level of mastery. The questions on the test are only a small sample of the many questions that could have been devised to measure the objectives of the test. The student took the test at a particular time, on a particular day, under a particular set of conditions. If we were to construct another test designed to measure the same objectives but containing different questions, or if the student took the same test a different time, on a different day, under different conditions, the student's score would probably be different. The effect of these factors will often be large enough so that a few students you would ordinarily consider masters of the objectives will get test scores that are lower than some students you would ordinarily consider non-masters. The result is that no matter where you set the cutoff score, you will make some errors in deciding which students have mastered the objectives and which have not.

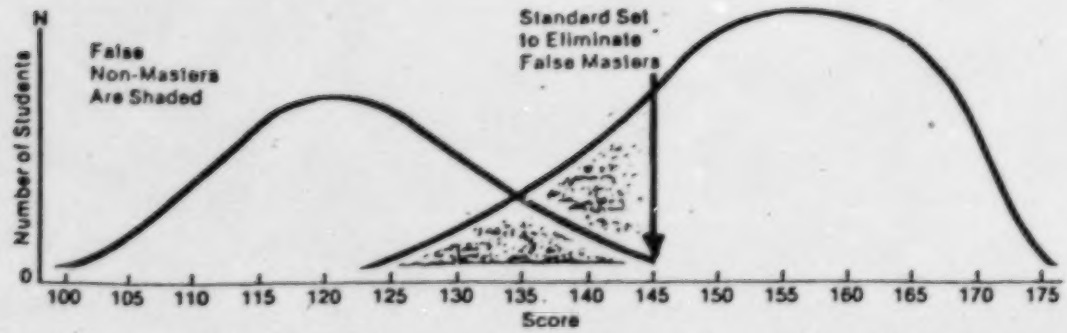
Two Types of Errors of Classification

There are basically two types of errors you can make:

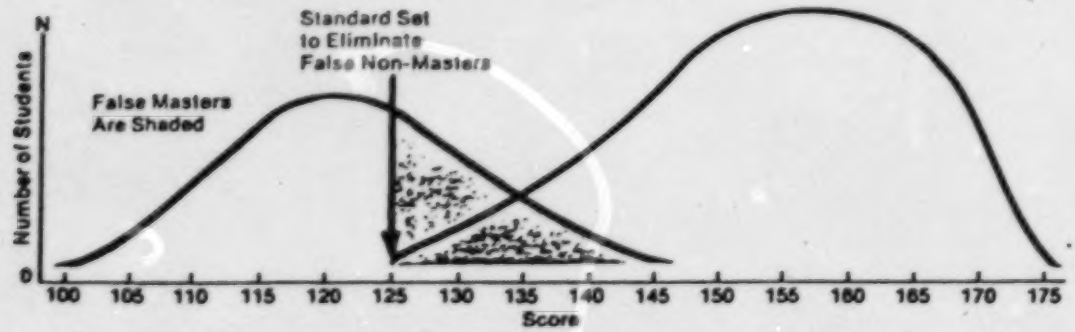
1. Classifying as a "master" a student who has not adequately mastered the objectives.
2. Classifying as a "non-master" a student who has adequately mastered the objectives.

We will refer to the first type of error as a "false master" and the second type as a "false non-master." By raising the standard, you can eliminate the first type of error (false master), but you will make a larger number of errors of the second type (false non-master).

Exhibit 3



By lowering the standard, you can eliminate false non-masters, but you will have a larger number of false masters.



By setting the standard where the curves cross, you can minimize the overall number of errors of classification.

If you are using the contrasting-groups method to set the standard, you can determine how many errors of each type you can expect to make and set the standard at the point at which the results are least harmful. But if you are using some other method, you cannot predict how many errors of each type you are likely to make. In general, though, if there are many students with scores close to the standard, there are likely to be a number of incorrect classifications. If one type of error is much more serious than the other, you may choose to adjust the standard so as to minimize that type of error, even at the cost of making more errors of the other type. For example, we used the data presented on page 29 to set the cutting score at 58. At this score point, there are 22 false

non-masters--students rated as masters who received scores lower than 58. By lowering the standard to score 49, we would reduce the number of false non-masters from 22 students to 3 students. By making this change, however, we would increase the number of false masters (students rated as non-masters who scored at or above standard) from 43 to 80.

One possibility to help reduce the errors of classification is to establish an "uncertain" category. For example, on the curves shown above, students scoring above 145 are extremely likely to be masters. Students scoring below 125 are extremely likely to be non-masters. Students scoring between 125 and 145 may be classified as "uncertain" and be further evaluated to help find their true status.

Please note that the contrasting-groups method does not cause more errors of classification than other methods do. But it is only with the contrasting-groups method that the kinds of errors of classification that you are likely to make become visible. You can move the standard to different points and estimate the kinds of misclassifications that are likely to result.

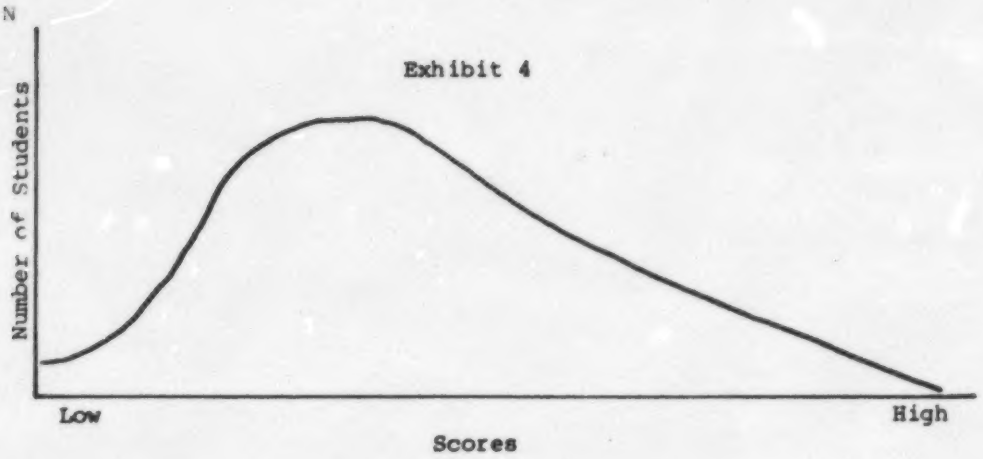
Because errors of classification are bound to occur, a cutting score should be viewed as an estimate until it is verified as reasonable. Several options exist for districts to further evaluate the cutting score or standard.

Determining Whether the Cutting Score is Reasonable

The cutting score should be reasonable with respect to both the impact it will have on the students and the district's expectations for educational quality. To determine the impact of a cutting score, it is best to plot the number of students receiving different scores on the test and examine the data for irregularities.

- Note where a chance score would fall if the test was multiple-choice. In most cases, the cutting score should be higher than a score obtained by chance guessing.
- Look at the distribution of scores for all students combined. Certain shapes can be expected, but others may indicate potential problems for a basic skills test. For instance, positively skewed scores (Exhibit 4)

Exhibit 4



suggest that the test may not be functioning as intended. If unexpected score distributions occur, further item and subgroup analyses may help pinpoint the problem.

- When using the contrasting-groups method, a contingency table can be set up to see the magnitude of potential errors of classification (Exhibit 5). To the extent that the percentage of possible errors is relatively small (assuming the teacher judgments are reliable), the more confidence one can have in the placement of the cutting score.

Exhibit 5

Teacher Judgments

(n = number of students)

	Judged Masters	Judged Non-Masters	TOTAL
Above the cutting score	Accurately predicted Masters n=900 (90%)	Possible errors: False-Masters n=100 (33%)	1000
Below the cutting score	Possible errors: False Non-Masters n=100 (10%)	Accurately predicted Non-Masters n=200 (67%)	300
TOTAL	1000	300	1300

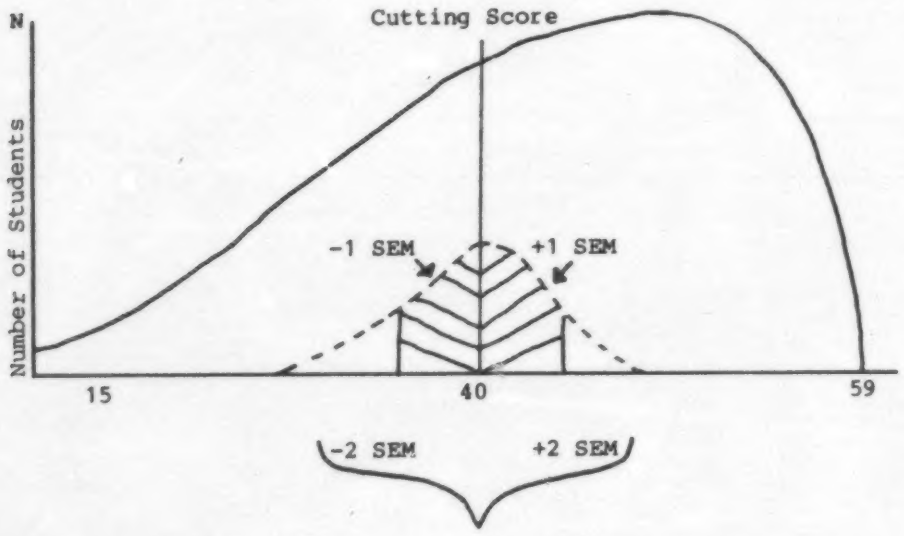
Adjusting the Cutting Score

Districts may want to consider the standard error of measurement (SEM) or standard deviation in evaluating students' scores[1]. The SEM should be reported in the technical manual. This is the overall error for the total test. It is also possible to calculate specific SEMs for various proposed cutting scores. The numerical value of this specific SEM will vary depending upon where the cutting score is set. In calculating this specific SEM, we are treating the cutting score (e.g., 40) as if it were a hypothetical student's true score. We are asking the question, what range of observed scores would be likely to cluster around a given true score (i.e., the cutting score of 40).

While the first edition of this manual (December 1979, p. 45) presented examples using the SEM for the total test, it is technically better to calculate a specific SEM for the particular cutting score you establish.

To see how the calculations are done and to assess the impact of the SEM, consider the hypothetical distribution in Exhibit 6.[2]

Exhibit 6



[1] The standard error of measurement is the standard deviation of actual test scores obtained (observed scores) around the true score (a hypothetical test score that is entirely free of error).

[2] The following procedures are not technically precise because in actuality the SEM represents a range of observed scores around a true score. However, if you follow these procedures for practical purposes you will be right more often than you will be wrong.

Our first task is to calculate a specific SEM for the proposed cutting score. We use the following formula:

$$\sqrt{\frac{X\ (N-X)}{N-1}}$$

N = the total number of items in the test
X = the cutting score selected

Suppose, for example, that we are setting a standard for the state mathematics test, consisting of 59 items, and that the proposed cutting score of 40 was calculated by one of the standards-setting methods. Using the above formula, the SEM for score 40 equals 3.62.[3] Using 1 SEM, we can be 68 percent confident that a student who actually obtained a score of 40 (the lowest possible score he or she could receive and still be considered minimally competent) would have a true score within the range of 36 (36.38) and 44 (43.62). The range of 36 to 44 represents the actual score plus and minus one standard error (40 ± 3.62).

If we wanted to be more confident in establishing a score band that encompassed the true score, we could extend the range to plus and minus two standard errors. By doing this, we would be 95 percent sure that an actual score of 40 represented a true score between 33 and 47 (40 ± 7.24).

Having established these score bands, there are numerous choices for adjusting the standard. You could:

1. Adjust your standard from 40 to 44 or 47, thus increasing the certainty of providing supplementary instruction for those who may need it.

2. Adjust your standard from 40 to 36 or 33, and by extending the acceptance range downward, reduce the risk of misclassifying students who have actually mastered the material.

In making either of the above kinds of adjustments, remember the trade-offs. By decreasing the probability that true masters are misclassified as needing

---

[3] $\sqrt{\frac{40\ (59-40)}{58}} = 3.62$

supplemental instruction, the probability of misclassifying remedial students as non-masters increases, and vice versa.

An Alternative to a Single Cutting Score

Even if we felt certain about the placement of the cutting score on a valid and reliable test, students' scores on the test would not be a perfect indication of their true level of mastery. There would still be error from other sources. For example, if the students took the test at a different time or under different conditions, their scores would probably be different. Consequently, there may be little or no real difference between two students when one scores just above and the other just below the cutting score (on the same test administration). Therefore, some additional safeguards may be useful.

One alternative to consider is the use of a band of scores around the cutting score or standard. The band could be defined as the standard plus and minus one or two standard errors of measurement. Students whose scores fall in this range might be examined more closely to determine whether or not they need supplemental instruction. Further examination might be based on a combination of test scores and teacher judgement.

Let's assume, for example, that a hypothetical school district sets the cutting score on the state reading test at score 40. The state reading test consists of 55 items and has a standard error of 3.33 for cutting score 40.[4] Using a score band of 40 $\pm$ 1 SEM, the district establishes a score range of 37 to 43 items. Students whose reading test scores fall within this range will be re-evaluated by teachers.[5]

---

[4] $\sqrt{\frac{40\ (55-40)}{54}} = 3.33$

[5] The re-evaluation process must be systematic; two sets of decision rules are needed. First, districts must decide what additional evidence (for example, other test scores, observations, etc.) teachers should use in re-evaluating students. Second, districts must determine how this additional information is to be applied in re-evaluating students. The information must be applied in a consistent manner so that all re-evaluated students are treated in the same way.

The district adopts the following decision rules to apply to students with test scores ranging from 37 to 43.

- Students with reading scores ranging from 37 to 43 will be systematically evaluated by teachers.
- Students who are rated above standard by teachers and have reading score 40 or above will be judged masters.
- Students who are rated below standard by teachers and score below 40 will be judged non-masters.
- Students whose teacher ratings and test scores are at variance will be retested.[6]

These decision rules are summarized below:

		Test Score	
		40 or Above	Below 40
Teacher Assessment	Above Standard	Masters	Retest
	Below Standard	Retest	Non-Masters

Using Multiple Methods to Set Cutting Scores

The concurrent validity of a cutting score in a given skill area can be demonstrated by reaching the same score by two independent methods. For instance, a district could apply one of the item-review methods, perhaps Nedelsky, and also a performance-based method, such as the contrasting-groups method: the closer the agreement of the cutting scores derived from each method, the more confidence one would have in the ultimate decisions. It is highly unlikely that two methods will yield the same cutting score: differences are to be expected. A difference of a few points (for example, within two standard errors of measurement) is reasonable. In such a case, several options exist for determining the final cutting score. If one has equal confidence in how both procedures were carried out, one might simply take

[6]When such retesting occurs, the original score and the retest score should be averaged to obtain a final score. This final averaged score should be the one used in determining the number and percentage of retested students who are considered above and below minimal standard.

the arithmetic average of the two scores. If there is more confidence in one procedure than the other, the cutting score can be weighted toward the more precise one. Or, the difference between the two scores can be used to define the interval or band discussed above and be treated as a questionable range requiring further review of student performance.

Monitoring Standards

For reasons already discussed, cutting scores obtained by any method should be viewed as tentative estimates. In addition to the controllable and uncontrollable sources of errors, many districts will be engaging in standards-setting for the first time and will make honest mistakes. Furthermore, the conditions that lead to judgments this year may be quite different in another year or two. Student populations may change, curricula may change, and instructional improvements at lower grades may lead to more proficient students entering at the higher grades and higher expectations by teachers or other judges.

For these reasons, standards should be reviewed annually. Ideally, predictive validity studies, even on an informal level, should be considered for this purpose. If feasible, for example, follow-up surveys of students who have graduated might be conducted to demonstrate the relationship between "success" in the adult world and performance on basic skills measures.

If such studies do not seem reasonable, consider other information at your disposal. You might interview a representative sample of students in grades 8 - 12 to determine the percentage that believe they will have difficulty finding work or continuing their education because they have problems in reading, writing, listening, or mathematics. You could use school records to help determine the proportion of graduated students that are neither employed or in postsecondary school. You could interview a sample of parents to determine the percentage that believe their children are having difficulty because they lack the basic skills.

There should be some congruence between the percentage of students who fall below standards that are set and the information obtained by the methods described above.

VI. REGULATIONS GOVERNING STANDARDS-SETTING AND REPORTING

How Does Public Participation Fit into the Standards-Setting Process?

The Regulations for Implementation of the Policy on Basic Skills Improvement require the genuine participation in the setting of standards of parents of public school students, Title I parents, employers, students (at the secondary level), and representatives of the general public as well as teachers and administrators.

The methods of standards-setting described in this manual rely heavily on the professional judgment of teachers. (In some methods, members of various other groups can also serve as judges.) These methods, however, can be employed in a standards-setting process that provides for public participation. The overall standards-setting process involves several steps.

- The formulation of proposed standards based on professional judgment
- A review of proposed standards by some process (committees, public hearings, surveys, etc.) which involves community participation
- The recommendation of standards based on professional judgment and community reactions provided for school committee review

This manual describes several methods of standards-setting. If possible, it is best to use more than one method, enabling staff and the public to review a range of defensible standards.

In reviewing the proposed standards, the community would presumably give consideration to one or more of the following:

- The overall impact of the standard in terms of the number of students whose scores are above and below the cutting score
- Resources the community has available to work with below standard students
- The advantages and disadvantages of setting the highest possible defensible standard to allow the greatest number to be considered for follow-up instruction

- The advantages and disadvantages of setting the lowest possible defensible standard to limit the number of students who might be labelled or unwittingly stigmatized because they scored below the minimal standard
- Local community needs and expectations
- The general ability level of the students in the district

As this list of issues indicates, there is ample opportunity for community involvement in the standards-setting process. The following chart summarizes some possibilities for public participation using the standards-setting methods described in this manual.

Suggestions for Public Participation

Methods	Defining minimally acceptable performance prior to standards-setting	Serving as judges - rating expected performance	Reviewing proposed standards	Comparing standards obtained using this method with standards derived from other methods
Nedelsky	X	X	X	X
Angoff	X	X	X	X
Borderline-Group	X		X	X
Contrasting-Groups	X		X	X

What Should be Reported?

By August 31, 1981, and no later than August 31 of each subsequent school year, school districts are required to report the number and percentage of students who have achieved, have not achieved, and have been exempted from achieving minimum standards in writing, reading, and mathematics.[6] For example, if you set your cutting score on the state mathematics test at 40, you would report the number and percentage of students who scored 40 or higher as above standard. Those scoring 39 or lower would be reported as below standard. If you adjusted your standard from 40 to 36 (40 - 1 SEM), you would report 36 as your cutting score.

This information must be reported for each grade level included in your district's Basic Skills Improvement Plan, and aggregated by student sex and race/ethnicity. If students are tested twice during a given school year, the final test results are the ones used in determining the number and percentage of students above and below standard. If you are using more than one test in a given skill area at a particular grade level (for example, subtests drawn from different forms), this should be noted on the reporting form. If you have questions about the use of more than one test, contact the Bureau of Research and Assessment. (A sample reporting form is included as Appendix G to this manual.)

In addition to reporting results directly to the state, districts must provide assurances that the Annual Report will be released and made generally available to the public, and that reasonable steps will be taken to ensure that the public is informed of the content and availability of the Annual Report.

---

[1] When listening and speaking are implemented, reporting on these skills will also be required.

APPENDIX A

Judge's Recording Form

Nedelsky Method

Test Title: _____

Judge's Name: _____

Question Number	Circle Number of Choices Identified	P	Question Number	Circle Number of Choices Identified	P
1	0 1 2 3		39	0 1 2 3	
2	0 1 2 3		40	0 1 2 3	
3	0 1 2 3		41	0 1 2 3	
4	0 1 2 3		42	0 1 2 3	
5	0 1 2 3		43	0 1 2 3	
6	0 1 2 3		44	0 1 2 3	
7	0 1 2 3		45	0 1 2 3	
8	0 1 2 3		46	0 1 2 3	
9	0 1 2 3		47	0 1 2 3	
10	0 1 2 3		48	0 1 2 3	
11	0 1 2 3		49	0 1 2 3	
12	0 1 2 3		50	0 1 2 3	
13	0 1 2 3		51	0 1 2 3	
14	0 1 2 3		52	0 1 2 3	
15	0 1 2 3		53	0 1 2 3	
16	0 1 2 3		54	0 1 2 3	
17	0 1 2 3		55	0 1 2 3	
18	0 1 2 3		56	0 1 2 3	
19	0 1 2 3		57	0 1 2 3	
20	0 1 2 3		58	0 1 2 3	
21	0 1 2 3		59	0 1 2 3	
22	0 1 2 3		60	0 1 2 3	
23	0 1 2 3		61	0 1 2 3	
24	0 1 2 3		62	0 1 2 3	
25	0 1 2 3		63	0 1 2 3	
26	0 1 2 3		64	0 1 2 3	
27	0 1 2 3		65	0 1 2 3	
28	0 1 2 3		66	0 1 2 3	
29	0 1 2 3		67	0 1 2 3	
30	0 1 2 3		68	0 1 2 3	
31	0 1 2 3		69	0 1 2 3	
32	0 1 2 3		70	0 1 2 3	
33	0 1 2 3		71	0 1 2 3	
34	0 1 2 3		72	0 1 2 3	
35	0 1 2 3		73	0 1 2 3	
36	0 1 2 3		74	0 1 2 3	
37	0 1 2 3		75	0 1 2 3	
38	0 1 2 3				

SUM ☐

APPENDIX B

Sample Judge's Recording Form

Angoff Method

Test Title: _____

Judge's Name: _____

Question Number	Estimated Probability	Question Number	Estimated Probability	Question Number	Estimated Probability
1	____	26	____	51	____
2	____	27	____	52	____
3	____	28	____	53	____
4	____	29	____	54	____
5	____	30	____	55	____
6		31		56	
7	____	32	____	57	____
8	____	33	____	58	____
9	____	34	____	59	____
10	____	35	____	60	____
11	____	36	____	61	____
12	____	37	____	62	____
13	____	38	____	63	____
14	____	39	____	64	____
15	____	40	____	65	____
16	____	41	____	66	____
17	____	42	____	67	____
18	____	43	____	68	____
19	____	44	____	69	____
20	____	45	____	70	____
21	____	46	____	71	____
22	____	47	____	72	____
23	____	48	____	73	____
24	____	49	____	74	____
25	____	50	____	75	____

SUM ☐

APPENDIX C

Borderline-Group

Sample Instructions for Rating Students' Reading Proficiency[1]

Directions for the Teacher

The Massachusetts Policy on Basic Skills Improvement requires local districts to establish minimum performance standards in reading at three grade levels. This district has chosen the grade level at which you are teaching as one of the three.

You are asked to identify students whose reading achievement, in your judgment, is not quite adequate, but not quite inadequate. These are students whose performances are so close to the borderline between mastery and non-mastery that it is difficult to say whether they need supplemental reading instruction.

Be very careful that your rating of "borderline" is based on achievement that you have observed. Do not rate students whom you do not know well enough to judge, and do not let students' attitudes or behavior influence your judgment.

Using the class roster which accompanies these directions , place a "B" next to the name of each student whose reading performance is not quite adequate, but not quite inadequate.

THESE RATINGS WILL BE COMPLETELY CONFIDENTIAL. NO RECORD WILL BE MADE OF ANY INDIVIDUAL RATING. THE RATINGS WILL BE USED SOLELY FOR ANALYZING GROUP DATA TO SET MINIMUM STANDARDS.

[1] These sample instructions pertain to reading. If you use this form for other skill areas, do not forget to change references to mathematics, writing, etc.

APPENDIX D

Contrasting-Groups

Sample Instructions for Rating Students' Reading Proficiency[1]

Directions for the Teacher

The Massachusetts Policy on Basic Skills Improvement requires local districts to establish minimum performance standards in reading at three grade levels. This district has chosen the grade level at which you are teaching as one of the three.

You are asked to rate each of your students on whether he or she, in your judgment, meets minimum basic skills standards in reading. The basic skill reading objectives for this grade accompany these directions.

Please rate each student on the class roster list which accompanies these directions. Use the following ratings and definitions.

M = This student is definitely above minimal standards in reading and does not need follow-up instruction.

S = This student is definitely below minimal standards in reading and needs supplementary instruction.

UK = I do not have enough information to judge this student's reading proficiency (studen is newly enrolled or not taught by me on a regular basis).

or

This student's reading skills are not clearly above or below minimal standard - they are at the borderline.

THESE RATINGS WILL BE COMPLETELY CONFIDENTIAL. NO RECORD WILL BE MADE OF ANY INDIVIDUAL RATING. THE RATINGS WILL BE USED SOLELY FOR ANALYZING GROUP DATA TO SET MINIMUM STANDARDS.

[1] These sample instructions pertain to reading. If you use this form for other skill areas, do not forget to change references to mathematics, writing, etc.

APPENDIX E

Contrasting-Groups

Sample Teacher Rating Form

School _____ Grade level _____

PLEASE LIST THE NAMES OF STUDENTS YOU NOW TEACH IN _____.

THEN RATE EACH STUDENT'S BASIC SKILLS PROFICIENCY IN _____.

USE THE DEFINITIONS BELOW.

M = This student is definitely <u>above</u> minimal standard and does not need follow-up instruction.

S = This student is definitely <u>below</u> minimal standard and needs supplementary instruction.

UK = I do not have enough information to judge this student's proficiency, <u>or</u> this student's skills are at the <u>borderline</u> between minimally acceptable and unacceptable.

Student Name	Teacher Rating	Test Score (Leave this column blank for later use)

APPENDIX F

THE CONTRASTING-GROUPS METHOD PARADOX

I. The Paradox

A desirable feature of a standards-setting method is that if standard setters in two different locations have the same standard in mind and use the same process, the expected values of standards they set should be equal.[1] In spite of all the advantages inherent in the contrasting-groups method, it does not have this characteristic. If an identical process is followed at two different sites, and the raters at both sites judge mastery and non-mastery the same way, then the site with a greater proportion of non-mastery students will be expected to establish a higher standard than the other site.

This problem can be shown readily. Consider the example shown in Table 1. A 10-item test has been administered to 360 masters and 120 non-masters. The smallest number of classification errors are made if the standard is set at a score of 5. Using this standard, a total of 33 classification errors would be made; 9 masters would fail this standard while 24 non-masters would pass. Using any other standard would result in a larger number of classification errors.

Now, suppose the identical process were conducted at another site. For purposes of our example, the only difference that will exist at this new site is that there are twice as many non-masters as there were at the first site. If that were the case, the data in Table 1 would be modified to look like the results in Table 2. The only difference between the two tables is that twice

---

[1] The standards set at the two different locations will not be identical every time, because there always will be some variability in the process. That is, at each site, if the same process were used several times, the standard would be a little higher sometimes, and a little lower other times. But over the long run, the average (or expected value) standard set at both locations should be identical.

as many non-masters have achieved each raw score on the test. Thus, where only three non-masters had a score of 0 at the first site, six have that score at the second site; nine had a score of 1 at the first site and eighteen at the second; and so on. Since the distributions of masters and non-masters are identical at both sites, we know that the way masters and non-masters are being classified must be identical at both sites.

TABLE 1.

A Hypothetical Set of Standard-Setting Data

	Mastery Students		Non-Mastery Students		
Score	Frequency	# Below Score	Frequency	# At or Above Score	Total # of Classification Errors
10	36	324	0	0	324
9	60	264	0	0	264
8	108	156	3	3	159
7	81	75	3	6	81
6	42	33	3	9	42
5	24	9	15	24	33
4	6	3	21	45	48
3	3	0	39	84	84
2	0	0	24	108	108
1	0	0	9	117	117
0	0	0	3	120	120

TABLE 2

The Same Hypothetical Data as in Table 1,
But with the Number of Non-Mastery Students Doubled

	Mastery Students		Non-Mastery Students		
Score	Frequency	# Below Score	Frequency	# At or Above Score	Total # of Classification Errors
10	36	324	0	0	324
9	60	264	0	0	264
8	108	156	6	6	162
7	81	75	6	12	87
6	42	33	6	18	51
5	24	9	30	48	57
4	6	3	42	90	93
3	3	0	78	168	168
2	0	0	48	216	216
1	0	0	18	234	234
0	0	0	6	240	240

The difference between these two examples is that, whereas the standard at the first site was a score of 5, the standard at the second site will be 6. It is at a score of 6 that the fewest total number of classification errors will be made. Thirty-three masters would fail the standard, and eighteen non-masters would pass.

This now is the paradox. While standard setters at both sites have used the same process, the standard at the second site is higher. The difference in the two standards is solely a function of the number of non-masters at each site and has nothing to do with the standards that the standard setters have in mind. And, as mentioned earlier, the paradox becomes even more critical because a higher standard always will be set at the location with more non-masters. Thus, a site that has a higher proportion of non-masters than another will compound its problems by setting a higher standard.

II. A Proposed Solution

One way to solve the paradox noted is to use proportions for all computations rather than raw frequencies. If proportions were used, all locations that had the same standards in mind, no matter how many non-masters they had, would develop the same table for choosing a standard.

Let's go back to the two tables we used to demonstrate that the paradox exists and show how the use of proportions eliminates the paradox. Table 3 is identical to Table 1, except that, in Table 3, columns of proportions have been computed by dividing all the counts of masters by the total number of masters. For example, there were 324 masters who had a score below 10. This is 324/360 of the masters, or .90. The proportions of non-masters were developed in a similar manner. For example, there were 3 non-masters who had a score of 8 or above. This is 3/120, or .03, of the non-masters. Note that the site with twice as many non-masters would develop an identical column of proportions. For example, while they had 6 non-masters with a score of 8 or above, this was 6/240 of their non-masters, or .03.

By using proportions, both sites would set the same standard. Both would choose a score of 6 as their standard, because the sum of proportion

of errors would be smallest for that score. Both would observe that .09 of their masters would fail that standard, and .08 of their non-masters would pass it.

However, by using unweighted proportions of masters and non-masters at both sites, we have made the data look just as it would if at both sites there were an equal number of masters and non-masters. This is equivalent to assuming that 50 percent of the students were below the standard. We earlier showed that the higher the number of non-masters relative to masters, the higher the standard that will be set. Consequently, the procedure used to

TABLE 3

The Same Hypothetical Data as in Table 1,
But with the Numbers Converted to Proportions

	Mastery Students			Non-Mastery Students			
Score	Freq.	# Below Score	Proportion of Masters Below Score	Freq.	# At or Above Score	Proportion of Non-Masters Above Score	Sum of Proportion of Errors
10	36	324	.90	0	0	.00	.90
9	60	264	.73	0	0	.00	.73
8	108	156	.43	3	3	.03	.46
7	81	75	.21	3	6	.05	.26
6	42	33	.09	3	9	.08	.17
5	24	9	.03	15	24	.20	.23
4	6	3	.01	21	45	.38	.39
3	3	0	.00	39	84	.70	.70
2	0	0	.00	24	108	.90	.90
1	0	0	.00	9	117	.98	.98
0	0	0	.00	3	120	1.00	1.00

develop Table 3 will result in standards that are too high, since, in most districts, far fewer than 50 percent of the students will be non-masters.

A more reasonable standard would be set if we assumed that there are more masters than non-masters; i.e., if the proportion of masters were multiplied by some standard amount. An arbitrary, but somewhat justifiable, multiplier of 10 could be used statewide. This would be equivalent to assuming that 9 percent of the students statewide were non-masters, and the costs of misclassifying masters were equal to the costs of misclassifying non-masters. These two assumptions seem reasonable, but one could readily argue for another value to be used for the multiplier. The point is that one should be chosen and used statewide by all districts that wanted to standardize their procedures.

The implementation of this final adjustment is shown in Table 4. This table is identical to Table 3, except that now "Proportion of Masters Below Score" has been multiplied by 10 for each score before being added to "Proportion of Non-Masters at or above Score." The standard selected now would be a score of 5, since the sum of errors is lowest (.45) for that score.

TABLE 4

The Same Hypothetical Data as in Table 1-3,
But with Proportion of Masters Weighted 10 Times

	Mastery Students				Non-Mastery Students			
Score	Freq.	# Below Score	Proportion of Masters Below Score	10 Times Proportion of Masters Below Score	Freq.	# At or Above Score	Proportion of Non-Masters Above Score	Sum of Proportion of Errors
10	36	324	.900	9.00	0	0	.00	9.00
9	60	264	.733	7.33	0	0	.00	7.33
8	108	156	.433	4.33	3	3	.03	4.36
7	81	75	.208	2.08	3	6	.05	2.13
6	42	33	.092	.92	3	9	.08	1.00
5	24	9	.025	.25	15	24	.20	.45
4	6	3	.008	.08	21	45	.38	.46
3	3	0	.000	.00	39	84	.70	.70
2	0	0	.000	.00	24	108	.90	.90
1	0	0	.000	.00	9	117	.98	.98
0	0	0	.000	.00	3	120	1.00	1.00

COMMONWEALTH OF MASSACHUSETTS
DEPARTMENT OF EDUCATION
BASIC SKILLS IMPROVEMENT POLICY

1980-81 ANNUAL REPORT ON BASIC SKILLS IMPROVEMENT PROGRAMS

System _____

System Code _____

Section 1 - Student Achievement of Minimum Standards

C. Secondary (7-12)[1]

(1) Skill (circle one): Reading Writing Mathematics

(2) Grade Level Reported (circle one): 7 8 9

(3) Evaluation Instrument(s) _____

_____

(4) Description of Minimum Standard _____

_____

(5) Students Evaluated for Achievement of Minimum Standards	All Students		Sex				Race/Ethnicity							
			F		M		B		H		W		O	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%
(a) Students Achieving Minimum Standards														
(b) Students Not Achieving Minimum Standards														
(c) Total Evaluated		100%		100%		100%		100%		100%		100%		100%

(6) Students Exempted/Not Evaluated	All Students	Sex		Race/Ethnicity			
		F	M	B	H	W	O
	#	#	#	#	#	#	#
(a) Special Education Exemption							
(b) Limited English Ability Exemption							
(c) Other							
(d) Total Exempted/Not Evaluated							

APPENDIX G

This manual was developed by:

Eileen Peters

Educational Testing Service
Northeastern Regional Office
111 Washington Street
Brookline, Massachusetts 02146